Industry competitions

Phil Culliton PHILCULLITON@GOOGLE.COM
Kaggle, Google, USA

Wei-Wei Tu TUWEIWEI@4PARADIGM.COM
The 4th Paradigm, China

Evelyne Viegas EVELYNEV@MICROSOFT.COM
Microsoft, USA

Mouadh Yagoubi MOUADH.YAGOUBI@TII.AE
Technology Innovation Institute (TII)

Hieu Khuong THANH-GIA-HIEU.KHUONG@UNIVERSITE-PARIS-SACLAY.FR
Université Paris-Saclay, France

Reviewed on OpenReview: *https://openreview.net/forum?id=XXXX*

Abstract

Industry-driven AI competitions are used by companies to push technical progress, evaluate solutions, and identify talent. By presenting well-defined real-world datasets and problems, these contests enable organizations to benchmark algorithms, discover new approaches, and inform strategic decisions about product development and resource allocation. For participants, they provide an opportunity to test skills, build a portfolio, and gain recognition. This chapter describes the role of such competitions in advancing the state of the art, highlights their value for both hiring and innovation, and outlines their relationship to benchmarks in AI research.

Keywords: industry, hiring, talent discovery, real-world datasets

With the rapid development of AI technology, companies pay increasingly more attention to AI applications and improvements. Organizing a machine learning competition not only furthers the development of specific machine learning workloads, but benefits the AI industry at large. For organizers, holding a successful AI contest is a leading way to surface the best algorithm and model for a new problem or to source innovative ideas to improve an existing workflow. For example, a company can decide how to allocate their team's data scientists: improving an existing model or developing a new one. A machine learning contest can help understand the current state-of-the-art in a problem space and help guide future workload decision-making resources in their science team.

Enterprises can also find and hire talent through such contests. A skill-based contest can present an industry-specific problem and a dataset that can identify world-class practitioners for a particular problem. For participants, joining can help them learn and practice their ML skills with a real-world application, develop a work portfolio, test new techniques, and gain recognition in the industry. For the AI industry as a whole, a skill-based contest is a great way to investigate the costs and benefits of different algorithms and models. They identify the current frontier of techniques, establish and publish new benchmarks, and set the standard for AI research. More details about the contribution they make to the AI industry are discussed in Section 1.2.

In this chapter, we introduce AI competitions hosted by industry partners ("industry competitions"). The remainder of this chapter is organized as follows. In Section 1, we briefly introduce the history of industry AI competitions, including the early form and modern form of the competitions

and the progress driven by industry AI competitions. In Section 2, we analyze different features of competitions involving different areas of business. In Section 4, we draw a conclusion about the content of this chapter.

1 A review of industry AI competitions: past and present

In this section, we briefly introduce the history of industry competitions. This includes 3 parts: 1) the history of industry competitions,; 2) progress achieved through industry competitions; 3) the present state of industry competitions.

1.1 The history of industry AI competitions

As an efficient way to measure the performance of the AI algorithm, competitions have a long history in their industry application. In 1997, IBM organized a chess competition between their chess AI "Deep Blue" and the chess world champion Gary Kasparov. In this competition, Deep Blue beat Kasparov and showed the world the great power and potential of artificial intelligence. The same story repeated in 2014, as Google researchers trained their "AlphaGo" program to learn, play, and ultimately defeat the human champion Yi Sedol in the board game Go - considered by many the final frontier of AI to surpass humans in strategy games.

A competition is also an important way for researchers and engineers to search for the AI algorithm and model with the best performance. In 1997, a direct marketing competition for the optimization of lift curves was held in the first Knowledge Discovery and Data Mining (KDD) Cup. Through this competition, researchers and engineers were invited to create a model to predict who is most likely to donate to a charity. This competition started the KDD Cup series , one of the most famous annual series of AI competitions that continues today (ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD), 2025). Ever since, a competition has been regarded as one of the most important ways to compare different machine learning models and find the state-of-the-art model structures.

In 2007-2009, Netflix created three competitions that comprised "The Netflix Prize", in which participants were challenged to build a collaborative filtering algorithm to predict user ratings for films. This competition became famous for its grand prize: $1,000,000 to the winning team that could beat Netflix's own algorithm. The competition gained a large media presence, drawing significant attention not only to the challenge itself but also to the power of competitions to address machine learning problems. Encouraged by the Netflix Prize, companies began to more widely utilize AI competitions around 2010.

The success of challenges such as the Netflix Prize highlighted a broader trend of using competitions to solve complex problems. Government agencies started with initiatives such as the National Institute of Standards and Technology Face Recognition Vendor Test (Ngan and Grother, 2015), which provides ongoing evaluations of facial recognition algorithms , while the DARPA Grand Challenge (Behringer et al., 2004) for autonomous vehicles that significantly accelerated the development of self-driving cars. In the private and non-profit sectors, organizations like the XPRIZE Foundation host high-profile competitions aimed at achieving technological breakthroughs (Hossain and Kauranen, 2014), and consortiums such as MLCommons have emerged to create standardized benchmarks such as MLPerf for fair industry-wide comparisons (Mattson et al., 2020). Concurrently, the rise of platforms like Kaggle further democratized access to industry datasets and prob-

lems, creating a global community of participants and making skill-based contests a mainstream tool for talent discovery and innovation.

1.2 Progress achieved through industry competitions

Progress driven by AI competitions mainly encompasses three aspects: the best solutions, identifying talent, and establishing benchmarks.

An AI competition quickly exhausts the problem space of a particular data science problem. By maintaining a leaderboard, participants are able to see what is currently possible when tackling this particular problem, and can use that to improve and iterate on their own models. Discussion forums often provide tips, tricks, and techniques for approaching the problem, examining the data, optimizing hardware, or determining which architectures are best. This collaborative, and competitive, environment invites quick iteration at the beginning and a thorough examination of the problem space from all angles. Because of this, competitions often quickly exceed existing benchmarks established by researchers or held by the enterprise internally. In the case of the Netflix Prize, it only took 6 days for the first participating team to surpass the performance of "Cinematch", their internal algorithm. Less than a year later, the top four teams competing in the Netflix Prize presented their techniques at the KDD Cup that year.

By presenting a specific, real-world business problem as part of a competition, an enterprise can get direct feedback on a problem, source a plethora of techniques, and practically compare the performance of machine learning and AI algorithms in the same way they'd be used in their business.

Typically, by combining the best techniques identified from the solutions of the winning teams, the organizer of a competition can directly improve the industry-standard performance of an algorithm in production. As an example, General Electric improved their flight arrival prediction model by 40% through a Kaggle competition in 2012(Hamner et al., 2012), or RTE obtained a power dispatcher algorithm capable of avoiding up to 90% of remedial actions emitted by CO2 through a competition hosted on Codalab (Pavão et al., 2025, 2022).

Machine learning competitions also help companies recruit top talent to their teams. A company can host a competition where the primary motivation of a participant is an interview or a job opportunity. By positioning this recruitment as a competition, the company can present a machine learning problem their team is already working and evaluate the performance of participants compared to the work of their current employees. This allows the organizer to objectively evaluate the skills of a candidate through the most practical application- what they would be doing each day- and attract talent that is interested and skilled in that particular field. Companies like Meta, Walmart, Allstate, and the NFL have all used machine learning competitions to identify and hire top talent. However, a common pitfall of recruiting competitions is evaluating participants only on their success in a single competition. It is not uncommon for competitions to conclude with marginal differences in the accuracy of the top models. The difference in score between the first place and the 1000th place could easily be a fraction of a decimal place. When using a machine learning competition to recruit machine learning practitioners for employment, it is best to consider the full corpus of work of the participant, including their success in other competitions, their contributions to the community at large, and their ability to explain their results - important factors in a workplace setting.

Industry competitions also establish new benchmarks in their problem space, which create a positive brand association with the organizers, as they accelerate the development of the AI industry at large. Today, competition platforms such as AICrowd, DrivenData, Tianchi, CodaLab, Kaggle, and others have accumulated hundreds of thousands of datasets- many of which are large-scale datasets collected from real-world businesses.

1.3 Industry AI competitions today

With the extensive use of AI technology, more and more industry enterprises have sought the benefits of organizing AI competitions. There are mainly two ways to organize and take part in industry AI competitions: 1) on an AI competition platform; 2) through an AI academic conference.

Competition platforms such as Kaggle, Tianchi, and CodaLab provide lots of convenience to both the organizers and contestants: an established platform, submission validation, real-time automatic submission scoring, a live competition leaderboard, competition-specific discussion forums, community-based features, pre-programmed competition evaluation metrics, and most importantly, a community of participants. An existing platform means that an organizer does not need to create a competition platform from scratch, can leverage the expertise of external teams, and can tap into a thriving community of machine learning practitioners. These platforms have featured over one thousand competitions from business barons such as American Express, Google, the National Football League, and the Japan Exchange Group. The prize pool for competitions can often exceed $50,000, and even reach over $1,000,000!

Determining the total prize pool, often a requirement of hosting a competition on an existing platform, is an important factor in attracting the best participants and creating a positive brand sentiment for the enterprise. The competition prize pool has a direct impact on attracting participants to the competition. Although many participants have noted that the prize is not their primary motivation for participating, a larger prize not only draws significant attention but also demonstrates the level of respect and challenge given by the organizers to the problem. The fees and prizes for a competition often cost less than the annual salary of a single data scientist and generate thousands of hours of world-class research on the problem. A small prize can risk a negative brand reputation associated with the host getting high-quality work 'for pennies on the dollar.' However, the inverse is also true. A large prize serves as a positive outreach tool for the machine learning community by demonstrating that organizers respect both the challenge the problem presents and the effort invested by participants.

Many industry enterprises may also choose to hold AI competitions through an AI academic conference. For example, 4paradigm organized an autograph competition through the 2020 KDD Cup with ChaLearn, Stanford University, and Google (Xu et al., 2022). This competition provided a total prize of 33,500 dollars and received 2269 submissions from 237 participants. A competition associated with an academic conference can generate a greater reputation and has the potential to attract greater visibility from researchers.

Competitions today are much more widely accessible than they have been previously, and cover a broader array of topics. Compared to competitions at their inception, AI competitions have become a common step in a practitioner's data science education, and many individuals even use competitions to prepare for a career change. Competition datasets can be both simple and approachable for a beginner, as well as larger, more complex, and requiring extensive domain-specific knowledge. Below are the summary data of some famous computer vision competitions held through top con-

ferences in recent years as examples. These data are listed in Table 1, which clearly indicates the development and bloom of AI competitions in recent years.

Table 1: Summary of industry CV competitions at top conferences (2019–2021).

Competition name	Year	Conference	Organizer	Teams	Public dataset size
AutoCV	2019	IJCNN	4Paradigm	102	>230,000
AutoCV2	2020	ECML/PKDD	4Paradigm	34	—
3D+Texture Garment Reconstruction	2020	NeurIPS	ChaLearn	263	>2M frames
AliProducts Challenge: Large-scale Product Recognition	2021	CVPR	Alibaba Group	623	>20,000 scenes
Multi-camera Multiple People Tracking (MMP-Tracking) Challenge	2021	ICCV	Microsoft	42	>2.9M frames

1.4 The difference between academic competitions and industry competitions

Today, both academic and commercial institutions are organizing various competitions on a variety of platforms. There are many differences between industry competitions and academic competitions. The most common difference is the objective of competition. Academic competitions are usually held to solve generic problems of conceptual importance, speed up the research of a state-of-the-art model, promote a dataset, or test a new ML framework. Industry competitions, however, are usually held to solve a specific problem of practical economic importance, develop or improve upon an existing algorithm, enhance their brand with the ML community, or hire additional resources for their team.

These different goals can be seen in real competitions. An example of industry is the Netflix Prize (Bennett and Lanning, 2007), which offered $1,000,000 for a 10% improvement to its recommendation algorithm. At the time, reports estimated that a 1% reduction in customer cancellations driven by better recommendations was worth $10 million a year to the company (Netflix, Inc., 2006). In comparison, the academic ImageNet Large-Scale Visual Recognition Challenge 2012 challenge (Russakovsky et al., 2015) did not have a large cash prize. It aimed to solve the general problem of image classification, and its winning "AlexNet" model - with a top-5 error rate of 15.3% compared to the second best of 26.2% - led to transformative changes in the field of ML.

1.5 Competition as a tool to leverage collaboration between research communities

Industrial competitions can serve as a unique opportunity for research communities to collaborate in solving a specific problem. In many cases, the need to organize such a competition arises when an organization lacks the necessary knowledge or tools to address a particular challenge within their industry. This need has grown in recent years as AI techniques have penetrated various fields and have established themselves as innovative solutions to problems that were previously solved using traditional methods within the area of expertise of the entity facing the problem. To cope with the wave of AI, industry has activated other levers such as in-company training and the recruitment of young graduates who have completed hybrid programs combining expertise in a specific field with AI techniques. However, the rapid growth of AI has driven industries to innovate in the face of industrial challenges. Competitions have emerged as a powerful way to push the boundaries of what AI can contribute to various domains, such as physical sciences, chemistry, and biotechnologies.

From an industrial perspective, challenges can serve as a bridge between AI and various domains, fostering collaboration and making AI knowledge more accessible to teams. For example, in the field of physical simulation, a well-designed challenge focused on applying AI to solve a physical problem can strengthen collaboration between the AI and scientific computing communities, enabling the formation of hybrid teams to tackle the problem. This observation is confirmed by the results of recent competitions organized around industrial use cases, where the winning methods were developed based on a deep exploration of both machine learning techniques on one hand and scientific computing methods on the other. Beyond technical solutions themselves, competitions create unique ecosystems that foster knowledge transfer through multiple channels.

- Cross-disciplinary team formation: The complex nature of industrial challenges often requires teams with diverse experience. In physics-based AI competitions, successful teams frequently combine members with backgrounds in domain sciences, numerical methods, and machine learning.
- Public benchmarks and datasets: By providing standardized, high-quality datasets and evaluation protocols, competitions establish common ground for researchers from different fields. Open datasets for physical simulation continue to serve as benchmarks for multiple communities long after the competitions end.
- Shared code repositories: Most competitions encourage or require open-sourcing of winning solutions, creating a valuable library of approaches that synthesize techniques from multiple domains. These repositories serve as educational resources and starting points for future research.
- Community forums and discussions: During competitions, participant discussions often reveal interesting cross-pollination of ideas between fields. Questions posed by machine learning experts may highlight new perspectives for domain scientists, and vice versa.
- Post-competition publications: The academic outputs from competitions, like retrospective analyses, document the synthesis of approaches and serve as accessible entry points for researchers wanting to explore the intersection of different fields.

Competitions also accelerate the practical application of novel methods. Performance metrics comparing proposed methods with industrial baselines highlight the potential real-world impact of these approaches. In industrial settings where computational efficiency directly translates into cost savings or expanded design exploration, such competitions can motivate the adoption of new techniques that might otherwise remain purely academic.

In addition, challenges can facilitate communication and knowledge exchange among participants through complementary communication initiatives that often supplement these competitions. Workshops, tutorial sessions, and follow-up collaborations frequently emerge from the competition ecosystem, sustaining cross-disciplinary dialogue beyond the duration of the challenge itself. Looking ahead, the structure of the competitions themselves might evolve to further enhance collaboration. Multistage competitions with intermediate feedback, team-merging opportunities, or explicit incentives for cross-disciplinary approaches could further strengthen the bridge between communities. The success of recent scientific AI competitions suggests that carefully designed challenges can serve not only as technical benchmarks but also as catalysts for the formation of new research communities at the intersection of established fields.

2 Industry challenges and benchmarks of different AI technology

As established in the previous section, industry-led competitions are powerful tools to accelerate innovation, establish new benchmarks, and foster collaboration. However, the design of a competition that yields scientifically valid and operationally relevant outcomes is a nontrivial challenge. This section analyzes case studies to identify best practices for competition design, as well as common difficulties that arise, such as the challenge of robust problem formulation, the complexities of real-world data, and the need to ensure fairness for participants. Through these examples, we will demonstrate how the most successful competitions overcome these hurdles to establish a durable benchmark - a combination of a dataset, a standardized evaluation metric, and a protocol.

2.1 Challenges and benchmarks in machine learning

The following cases illustrate the evolution from the use of generic metrics that create a gap between competitive success and production viability to designing nuanced objectives that better reflect the complexities of industrial applications.

- **Netflix Prize**. The Netflix Prize has been held three times in 2007-2009. The challenge was to build a collaborative filtering algorithm to predict user film ratings, with a $1,000,000 prize to improve Netflix's baseline RMSE by 10%. The speed of progress demonstrated the power of industry competition: within the first week, a team had already surpassed the internal 'Cinematch' baseline, and the 1% threshold for the first Progress Prize was met just a week later. The contest attracted over 40,000 teams from 186 countries, culminating in a victory for "BellKor's Pragmatic Chaos", a team formed by merging several top researchers from the United States, Austria and Canada. While the competition successfully established a new state-of-the-art based on matrix factorization (Koren, 2009), the competition's reliance on a single metric, RMSE, created a significant gap between the winning solution and a production-ready system. The winning solution was a complex ensemble of over 100 models that was never fully implemented in production. Despite this, the competition's legacy was immense, through the creation of a foundational benchmark of a training set including 100,480,507 ratings that 480,189 users gave to 17,770 movies, and a qualifying test set containing over 2,817,131 ratings. This fueled a wave of scientific advancement, inspiring early deep learning approaches to recommendations (Salakhutdinov et al., 2007), scalable training methods (Hu et al., 2008), generalized models (Rendle, 2010), and new algorithms for implicit feedback (Rendle et al., 2012).

- **KDD Cup 2020 Challenges for Modern E-Commerce Platform: Debiasing**. In a departure from competitions focused solely on generic accuracy, the 2020 KDD Cup, hosted by Alibaba, exemplified a best practice in modern competition design: aligning the evaluation metric with a specific and nuanced business objective. The challenge addressed the common industrial difficulty of noisy and skewed long-tail distributions in recommendation systems, focusing on providing fair exposure for rarely seen products on an e-commerce platform, a problem where optimizing for overall accuracy often harms diversity and new sellers. To address this, the organizers created a sophisticated dual-metric evaluation. Submissions were judged not only on 'NDCG@50-full ' (overall performance), but also on 'NDCG@50-rare', a custom metric that explicitly rewarded models for their performance on underexposed items. This careful design created a productive tension, steering the 1,895 participating teams away

from simply optimizing for popular items and toward developing novel solutions using multistage architectures and graph neural networks. It also established a public benchmark for fairness and long-tail performance, including a training set and a qualifying test set contained over 1,000,000 clicks, 100,000 items, and 30,000 users across 10 periods including massive sales. This dataset has paved the way for a new wave of research into debiasing and balancing accuracy and fairness in recommendation systems (Chen et al., 2023; Huang et al., 2020; Xue et al., 2020).

2.2 Challenges and benchmarks in computer vision

Nowadays, deep learning models represented by convolutional neural networks (CNNs) often become the most popular approach in computer vision (CV) competitions and play important roles in industry applications such as face recognition, auto driving, auto retailing, etc. The success of these events hinges less on the winning algorithm and more on thoughtful competition design. Beyond the problem definition, organizers must contend with the realities of a dynamic world and a diverse participant base. The following cases demonstrate how design choices related to datasets, evaluation protocols, and participant resources can determine a competition's ability to test for generalization and robustness, ensure fairness, and create a lasting scientific impact.

- **Deepfake Detection Challenge**. A model's ability to generalize beyond its training data is paramount for real-world application. The Deepfake Detection Challenge (DFDC), hosted by Meta on Kaggle, serves as a seminal example of designing a competition to explicitly test for this. With a $1,000,000 prize, it drew 2,265 teams to develop models to detect AI-generated fake videos. The competition's most poignant lesson came from its dramatic "leaderboard shakeup". The final evaluation was conducted on a private test set that included manipulation techniques not present in the less general public training data. As a result, many top performing models on the public leaderboard plummeted in the final rankings; one model dropped from first place to 904th. This shakeup, driven by the discrepancy between training and private test sets, effectively measured true generalization rather than simple pattern matching, mirroring the industrial need for security systems that can withstand novel attacks. Furthermore, the DFDC addressed the critical issue of fairness in computing. Recognizing that training models on large video datasets requires immense GPU power, organizers collaborated with cloud providers to offer computing credits. This leveled the playing field, ensuring that the competition rewarded algorithmic innovation over access to hardware. Despite the issues with the training set's generality, the competition's legacy includes a public Deepfake detection datasets created by Meta, AWS, Microsoft, and academic experts, containing 124,000 videos featuring 8 facial modification algorithms (Dolhansky et al., 2020). This has fueled further research focused on more general benchmarks (Yan et al., 2023) and more novel detection methods (Gu et al., 2022; Oorloff et al., 2024).

- **Multi-camera Multiple People Tracking (MMP-Tracking) Challenge**. In contrast, the long-term impact of a competition can sometimes be the benchmark it creates rather than the solutions it identifies, a principle exemplified by the MMP-Tracking Challenge. This competition was organized by Microsoft, including two sub-tracks: 1) evaluating tracking result with top-down view, 2) evaluating tracking result from each camera. 42 teams consisting of researchers and engineers from industry companies participated in this competition. Al-

though the competition crowned winners, its most significant and lasting contribution was the release of the MMPTRACK dataset, which includes 5 environment, 23 cameras and 576 minute video data with over 2,979,000 frames (Han et al., 2023). As the first large-scale, high-quality public benchmark for this task, the dataset's comprehensive nature established a strong, reliable standard rather than causing a leaderboard shake-up. This dataset has become a cornerstone for subsequent research in the field, influencing scientific progress far more than the competition's final standings. Its release has fueled the development of further benchmarks (Woo et al., 2024) and novel scientific work, such as the Multi-Camera Tracking Transformer (Niculescu-Mizil et al., 2025).

These two cases offer complementary lessons. While the DFDC shows how a strategically designed private test set can pressure-test models for robustness, the MMP-Tracking Challenge illustrates that creating a foundational public dataset can be an even greater contribution, fostering years of innovation. Both highlight that because training CV models requires a significant computational load, providing resources is a best practice for ensuring fair and accessible competitions.

2.3 Challenges and benchmarks in natural language processing

Natural language processing (NLP) is an important basis of customer service that is significantly impacted by AI, such as voice assistance, question answering system, search engine optimization, etc. The most impactful competitions also teach meaningful lessons in problem formulation and evaluation design to address more nuanced real-world objectives. The following 2 examples about typical industry NLP competitions illustrate the importance of aligning evaluation metrics with specific, practical goals.

- **Diagnostic Questions: Predicting Student Responses and Measuring Question Quality**. Hosted by Microsoft, academic researchers, and the Eedi education provider at NeurIPS 2020, with 447 teams joining, this competition serves as an example in thoughtful problem formulation. Rather than posing a simple binary classification task on whether a student's answer is correct or incorrect, the challenge includes 4 different objectives: 1) predicting whether or not students will answer questions correctly; 2) predicting which multiple-choice answer students choose for each question; 3) devising a metric to measure the quality of the questions; 4) acquiring a limited set of answers from students in order to accurately predict student performance prediction on unseen questions. Submissions to tasks 1), 2) and 4) were evaluated by prediction accuracy, while task 3) was evaluated by a group of domain experts. Its multi-task objective required participants to predict not only correctness but also which specific multiple-choice distractor a student would select. This design choice represents the best practice of choosing an evaluation metric that was closely aligned with a real-world educational goal. The competition legacy is a dataset that includes more than 28,000 questions and 17,250,000 responses from 123,000 students (Wang et al., 2021). This has helped science progress in developing more techniques (Ghosh et al., 2021), including Reinforcement Learning (He-Yueya and Singla, 2021), and more explainable models (Berthon and van der Schaar, 2025).
- **Automated Language Processing for Text Categorization (AutoNLP)**. Held by 4paradigm, academic researchers, and Google at WAIC 2019 with a $7500 prize, this competition addressed the key industrial need for automation. As the exploitation of text categorization

nowadays is often relied on experienced human experts with in-depth domain knowledge in a labor-intensive trial-and-error manner, it is necessary to develop a single solution that could solve multiple, unseen text categorization problems, effectively a "meta-challenge" on AutoML for NLP. Participants received five training sets and five validation sets. The competition provided 3 baselines: SVM, CNN-rand (Kim, 2014) and CNN with fastText-pre-trained embedding models. To ensure a level playing field and enforce realistic resource constraints, all submissions were executed within a standardized computational environment with 1 NVIDIA Tesla P100 GPU each, hosted on the CodaLab platform (Pavao et al., 2023). By requiring submissions to be trained on 5 distinct datasets while evaluated on 5 different datasets, the competition established a benchmark to measure the generalizability of AutoML systems (Liu et al., 2020). The event advanced the state-of-the-art in resource-constrained automated deep learning methods for NLP (Tuggener et al., 2020; Luo et al., 2023; Gao et al., 2022), and paved the way for more AutoML competitions in other domains (Wang et al., 2020).

These two challenges highlight the maturation of NLP competitions. They demonstrate a shift away from a singular focus on leaderboard scores and toward a more sophisticated approach where the competition's very objectives, metrics, and evaluation protocols are crafted to solve a specific industrial or scientific problem. Whether it is by aligning metrics with real-world utility or by designing a framework to test automation, these competitions create lasting value by establishing not just benchmarks but best practices.

2.4 Challenges and benchmarks on LLMs

Rapid advancement of large language models (LLMs) has transformed the landscape of AI competition. Learning from the pitfalls of earlier competitions, the LLM community has embraced thoughtful competition design to address its unique, large-scale challenges. The most impactful competitions are now structured as scientific experiments, using carefully crafted constraints and evaluation protocols to isolate variables and produce clear, actionable insights. The following examples demonstrate how this mature approach to design yields conclusions directly relevant to critical industrial problems: the immense computational cost that limits access, and the profound security risks that hinder adoption.

- **NeurIPS 2023 LLM Efficiency Challenge**. The story of modern LLMs is one of scale, which creates a significant barrier to entry. With a prize pool of $30,000, the NeurIPS 2023 LLM Efficiency Challenge was conceived as a direct response to this accessibility crisis (Saroufim et al., 2025). Its design addressed the "fairness-in-computing" problem head-on. By imposing strict resource constraints, fine-tuning on a single consumer-grade GPU (RTX 4090 or A100) in under 24 hours, the organizers neutralized the variable of raw computing power. This best practice transformed the problem from a race for the largest model into a competition of ingenuity. The evaluation protocol was a customized version of Stanford's Holistic Evaluation of Language Models (HELM), assessing models on an open set of benchmarks (including MMLU, TruthfulQA, and GSM8k) and a hidden, closed set to test for generalization. The key lesson learned from the winning solutions was definitive: Under these constraints, meticulous data curation was far more impactful than novel model architectures. The top teams achieved state-of-the-art results by carefully filtering and mixing

public datasets like Open-Platypus and LIMA to create small but highly effective fine-tuning sets. The competition thus provided large-scale empirical evidence that for efficient LLM adaptation, data quality is a primary driver of performance, not raw scale.

- **2024 SaTML LLM Capture-the-Flag Competition**. As LLMs are integrated into applications, their security has become an paramount concern. Learning the lesson that static benchmarks often fail to capture real-world robustness, the 2024 SaTML LLM Capture-the-Flag (CTF) competition was designed as a live, dynamic, adversarial environment. The setup mimicked a real-world security scenario: 'defender' teams submitted 44 unique defenses to protect a secret "flag" embedded in models like 'gpt-3.5-turbo' and 'llama-2-70b-chat', while 'attackers' engaged in multi-turn conversations to try and extract it. This design was tested directly for adaptive, intelligent attacks rather than static vulnerabilities. The most significant lesson was the universal failure of the submitted defenses; every single one was bypassed at least once by adaptive attackers. This highlighted the insufficiency of simple filter-based security and the critical need for dynamic, multi-turn evaluation, as many successful attacks unfolded over several conversational turns. The competition contributed a lasting open source dataset of more than 137,000 multi-turn adversarial conversations (Debenedetti et al., 2024a). This benchmark provides an invaluable resource for the community to train and test more techniques (Khomsky et al., 2024) and inspired more research into the dynamic environment for safety benchmarks (Debenedetti et al., 2024b).

The focus on efficiency and security represents just two facets of the evolving LLM competition landscape. Research on LLM has therefore been rapidly evolving across multiple fronts, including how to evaluate them effectively (Weidinger et al., 2025), with platforms such as the LMSYS Chatbot Arena establishing new benchmarks based on human preference (Chiang et al., 2024), how to improve their efficiency (Wan et al., 2023), their training scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022), and how to combine or adapt them for various use cases(Hadi et al., 2023). In recent NeurIPS editions, several competitions have been introduced to address the high costs associated with training and deploying them: the LLM Merging Challenge explored how to combine existing models to create more powerful ones without additional training (Tam et al., 2024), while the Edge LLM Challenge (Liu et al., 2024) focused on developing efficient, optimized models capable of running on resource-constrained edge devices.

3 Industry AI challenges and benchmarks in different business area

In this section, we analyze the features of AI challenges and benchmarks of two specific business areas in which AI technology has been widely applied: financial technology ('fin-tech') and retail business.

3.1 Challenge and benchmarks of fin-tech

In the financial industry, AI technology is playing an increasingly important role- investment banks and hedge funds are seeking machine learning talent at the same levels, if not higher, than traditional finance or MBA professionals. Some of the largest hedge funds, like Two Sigma, have achieved their success with an AI-first approach, rather than a finance-first approach, as artificial intelligence has proven to demonstrate a significant impact on trading strategies. AI competitions, therefore, serve

as an important way for financial companies to acquire AI solutions and hire top talent. Below, we list 2 fin-tech competitions on Kaggle as examples.

- **JPX Tokyo Stock Exchange Prediction**. Organized by Japan Exchange (JPX) Group, the challenge was to evaluate the trend of the price of stocks. Participants were required to rank each stock active on a given day, including the top 200 and the bottom 200. Returns for a single day treated the top 200 stocks as purchased and the bottom 200 stocks as shorted. The weighted returns for the portfolio would be calculated assuming that the stocks were purchased the next day and sold the day after that. The goal of the participants was to maximize their weighted returns. JPX Group provided the stock price and quarterly earnings data of more than 2000 stocks. Over 2,000 teams joined the competition to compete for a 65,000 dollar prize for the top 10 teams (AkihiroSugiyama et al., 2022).

- **Zillow Prize: Zillow's home value prediction (Zestimate)**. Estimating the price and value of house is an important problem in the finance and estate industry. Organized by a Zillow estate enterprise, this competition focused on how to improve the prediction accuracy of the estate sale price. Zillow provided a full list of real estate properties and all the transactions data from January 1, 2016 to October 15, 2016 as the training set. The participants were then required to predict the price of the estate at 6 given time points. The performance of the predict model was measured by the log-error between the prediction and the real price. 3,770 teams joined this competition to compete for a total of 1,200,000 dollars prize (AndrewMartin et al., 2017).

3.2 Challenge and benchmarks of retail business

With the help of AI technology, the retail business is becoming more and more convenient and efficient. There are also many retail enterprises who choose to organize their own AI competitions. Here we introduce 2 examples as follows:

- **The fifth M Competition**. This was the fifth Makridakis competition (also known as M competitions) held by the University of Nicola in 2020, which included two tracks: the accuracy track and the uncertainty track (Makridakis et al., 2022). The challenge of the accuracy track was to estimate the unit sales of Walmart using machine learning technology, and the challenge of the uncertainty track is to estimate the uncertainty distribution. Both tracks use a common dataset that includes information about the dates on which the products are sold, the historical daily unit sales data per product and store, and the price of the products sold per store and date. The sale data contains 42,840 time series with more than 3,900 features. In each track, a total of 50,000 dollars prize was provided. 5,558 teams joined the accuracy track and 909 teams participated in the uncertainty track.

- **H&M Personalized Fashion Recommendations**. This competition was held by H&M Group on Kaggle in 2022 (Ling et al., 2022). The challenge is to develop a product recommendation model based on data from previous transactions, as well as from customer and product meta data. The training data includes 105,000 pictures and 4 tabular dataframes with a total size of 34.56GiB. Submissions were evaluated according to mean average precision. H&M Group provided a total of 50,000 dollars of reward for the top 6 teams and more than 2,900 teams joined this competition.

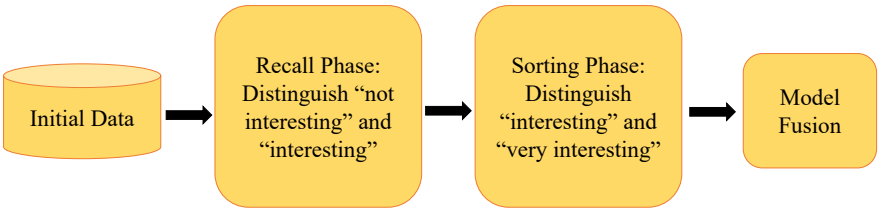

Figure 1: The process of solving product recommendation challenge

The challenges of retail business are mainly sales prediction and product recommendation. Sales prediction is a form of time series prediction problem and is often solved by models utilizing RNN, LSTM, etc. A challenge of product recommendation can often be solved by a fusion of a recall model and a sorting model, with a process shown in Fig.1.

4 Conclusion

In conclusion, industry AI competitions have proven to be a dynamic and influential mechanism in the AI ecosystem. These contests offer a unique convergence of industry-specific challenges and cutting-edge AI techniques, bringing together diverse participants and fostering practical innovation. For organizers, AI competitions serve as a vehicle to source the most effective algorithms, gather information from a wide range of approaches, and establish new industry benchmarks that drive long-term progress. In addition, they act as powerful talent pipelines connecting enterprises with world-class data scientists and engineers who excel in real-world problem solving.

For participants, contests provide an invaluable opportunity to apply theoretical knowledge in practical scenarios, to learn and refine skills, and to engage in a vibrant community of peers and mentors. Competing in these challenges not only strengthens technical competencies, but also builds a tangible portfolio that can accelerate professional growth and recognition.

However, as highlighted throughout this chapter, competitions are not without challenges and limitations. Overfitting to a leaderboard or relying on ensembling methods that are too complex for real-world deployment are common pitfalls. Organizers should be aware that the highest scoring solution may not be the most suitable for production environments. Instead, the real value of these contests often lies in the diversity of solutions and the innovative ideas they generate.

Looking ahead, the landscape of AI competitions continues to evolve. With the rapid development of generative AI and large language models, new frontiers of challenges and opportunities are emerging. Future competitions might increasingly focus on the efficiency, interpretability, and ethical implications of AI solutions, especially in mission-critical applications and areas of social impact. Multidisciplinary contests that encourage collaboration between domain experts and machine learning practitioners are likely to become more prominent, bridging the gap between theoretical advancements and real-world deployment.

We hope that this chapter has provided a comprehensive understanding of the multifaceted role of AI competitions in the industry. By exploring their history, highlighting successful case studies and analyzing their present and future impact, we aim to equip both competition organizers and participants with the insights and best practices necessary to maximize the value of these contests.

Ultimately, we believe that well-designed and thoughtfully executed AI competitions can play a pivotal role in shaping the future of AI and its transformative applications across industries.

References

ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD). Kdd cup: The annual data mining and knowledge discovery competition, 2025. URL https://kdd.org/kdd-cup.

AkihiroSugiyama, Chihiro Hio(Alpaca), Eiichiro Kaji, n onishi, s-meitoma JPX, Shun Takato, Shun Takato JPX, Sohier Dane, and Tomoya Kitayama(Alpaca). Jpx tokyo stock exchange prediction. https://kaggle.com/competitions/jpx-tokyo-stock-exchange-prediction, 2022. Kaggle.

AndrewMartin, Bin, Cat N, K Nielsen, Maggie, and Wendy Kan. Zillow prize: Zillow's home value prediction (zestimate). https://kaggle.com/competitions/zillow-prize-1, 2017. Kaggle.

Reinhold Behringer, Sundar Sundareswaran, Brian Gregory, Richard Elsley, Bob Addison, Wayne Guthmiller, Robert Daily, and David Bevly. The darpa grand challenge-development of an autonomous vehicle. In *IEEE Intelligent Vehicles Symposium, 2004*, pages 226–231. IEEE, 2004.

James Bennett and Stan Lanning. The netflix prize. 2007.

Antonin Berthon and Mihaela van der Schaar. Language bottleneck models: A framework for interpretable knowledge tracing and beyond. *arXiv preprint arXiv:2506.16982*, 2025.

Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 41(3):1–39, 2023.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

Edoardo Debenedetti, Javier Rando, Daniel Paleka, Silaghi Florin, Dragos Albastroiu, Niv Cohen, Yuval Lemberg, Reshmi Ghosh, Rui Wen, Ahmed Salem, et al. Dataset and lessons learned from the 2024 satml llm capture-the-flag competition. *Advances in Neural Information Processing Systems*, 37:36914–36937, 2024a.

Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for llm agents. *Advances in Neural Information Processing Systems*, 37:82895–82920, 2024b.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

Jiahui Gao, Hang Xu, Han Shi, Xiaozhe Ren, Philip LH Yu, Xiaodan Liang, Xin Jiang, and Zhenguo Li. Autobert-zero: Evolving bert backbone from scratch. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10663–10671, 2022.

Aritra Ghosh, Jay Raspat, and Andrew Lan. Option tracing: Beyond correctness analysis in knowledge tracing. In *International Conference on Artificial Intelligence in Education*, pages 137–149. Springer, 2021.

Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, and Lizhuang Ma. Delving into the local: Dynamic inconsistency learning for deepfake video detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 744–752, 2022.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*, 3, 2023.

Ben Hamner, Dyan Finkhousen, Feng (Fred) Xue, and joycenv. Ge flight quest. `https://kaggle.com/competitions/flight`, 2012. Kaggle.

Xiaotian Han, Quanzeng You, Chunyu Wang, Zhizheng Zhang, Peng Chu, Houdong Hu, Jiang Wang, and Zicheng Liu. Mmptrack: Large-scale densely annotated multi-camera multiple people tracking benchmark. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4860–4869, 2023.

Joy He-Yueya and Adish Singla. Quizzing policy using reinforcement learning for inferring the student knowledge state. In *14th International Conference on Educational Data Mining*, pages 533–539. educationaldatamining. org, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Mokter Hossain and Ilkka Kauranen. Competition-based innovation: the case of the x prize foundation. *Journal of Organization Design*, 3(3):46–52, 2014.

Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE international conference on data mining*, pages 263–272. Ieee, 2008.

Jianqiang Huang, Ke Hu, Mingjian Chen, Bohang Zheng, Xingyuan Tang, Tan Qu, Yi Qi, and Jun Lei. KDD Cup 2020 Debiasing: 1st Place Winning Solution. `https://github.com/aister2020/KDDCUP_2020_Debiasing_1st_Place`, 2020. Accessed: 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Daniil Khomsky, Narek Maloyan, and Bulat Nutfullin. Prompt injection attacks in defended systems. In *International Conference on Distributed Computer and Communication Networks*, pages 404–416. Springer, 2024.

Yoon Kim. Convolutional neural networks for sentence classification. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1181. URL https://aclanthology.org/D14-1181/.

Yehuda Koren. The bellkor solution to the netflix grand prize. *Netflix prize documentation*, 81 (2009):1–10, 2009.

Carlos García Ling, ElizabethHMGroup, FridaRim, inversion, Jaime Ferrando, Maggie, neuraloverflow, and xlsrln. H&m personalized fashion recommendations. https://kaggle.com/competitions/h-and-m-personalized-fashion-recommendations, 2022. Kaggle.

Shiwei Liu, Kai Han, Adriana Fernandez-Lopez, AJAY KUMAR JAISWAL, Zahra Atashgahi, Boqian Wu, Edoardo Ponti, Callie Hao, Rebekka Burkholz, Olga Saukh, et al. Edge-llms: Edge-device large language model competition. In *NeurIPS 2024 Competition Track*, 2024.

Zhengying Liu, Zhen Xu, Shangeth Rajaa, Meysam Madadi, Julio CS Jacques Junior, Sergio Escalera, Adrien Pavao, Sebastien Treguer, Wei-Wei Tu, and Isabelle Guyon. Towards automated deep learning: Analysis of the autodl challenge series 2019. In *NeurIPS 2019 Competition and Demonstration Track*, pages 242–252. PMLR, 2020.

Zhipeng Luo, Jiahui Wang, and Yihao Guo. DeepBlueAI at PragTag-2023:ensemble-based text classification approaches under limited data resources. In Milad Alshomary, Chung-Chi Chen, Smaranda Muresan, Joonsuk Park, and Julia Romberg, editors, *Proceedings of the 10th Workshop on Argument Mining*, pages 202–206, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.argmining-1.23. URL https://aclanthology.org/2023.argmining-1.23/.

Spyros Makridakis, Fotios Petropoulos, and Evangelos Spiliotis. The m5 competition: conclusions, 2022.

Peter Mattson, Vijay Janapa Reddi, Christine Cheng, Cody Coleman, Greg Diamos, David Kanter, Paulius Micikevicius, David Patterson, Guenther Schmuelling, Hanlin Tang, et al. Mlperf: An industry standard benchmark suite for machine learning performance. *IEEE Micro*, 40(2):8–16, 2020.

Netflix, Inc. 2006 annual report. Technical report, Netflix, Inc., Los Gatos, CA, 2006. URL http://q4live.s22.clientfiles.s3-website-us-east-1.amazonaws.com/959853165/files/doc_financials/annual_reports/NFLX.pdf. Accessed: 2025-08-30.

Mei Ngan and Patrick Grother. *Face Recognition Vendor Test (FRVT):*. US Department of Commerce, National Institute of Standards and Technology, 2015.

Alexandru Niculescu-Mizil, Deep Patel, and Iain Melvin. Mctr: Multi camera tracking transformer. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 874–884, 2025.

Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. Avff: Audio-visual feature fusion for video deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27102–27112, 2024.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Dinh-Tuan Tran, Xavier Baro, Hugo Jair Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. Codalab competitions: An open source platform to organize scientific challenges. *Journal of Machine Learning Research*, 24(198):1–6, 2023.

Adrien Pavão, Isabelle Guyon, Anne-Catherine Letournel, Xavier Baró, Hugo Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. Codalab competitions: An open source platform to organize scientific challenges. *Technical report*, 2022. URL `https://hal.inria.fr/hal-03629462v1`.

Adrien Pavão, Antoine Marot, Jules Sintes, Viktor Eriksson Möllerstedt, Laure Crochepierre, Karim Chaouache, Benjamin Donnot, Van Tuan Dang, and Isabelle Guyon. Ai challenge for safe and low carbon power grid operation. *Energy and AI*, page 100564, 2025. ISSN 2666-5468. doi: https://doi.org/10.1016/j.egyai.2025.100564. URL `https://www.sciencedirect.com/science/article/pii/S2666546825000965`.

Steffen Rendle. Factorization machines. In *2010 IEEE International conference on data mining*, pages 995–1000. IEEE, 2010.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on Machine learning*, pages 791–798, 2007.

Mark Saroufim, Yotam Perlitz, Leshem Choshen, Luca Antiga, Greg Bowyer, Christian Puhrsch, Driss Guessous, Supriya Rao, Geeta Chauhan, Ashvini Kumar, et al. Neurips 2023 llm efficiency fine-tuning competition. *arXiv preprint arXiv:2503.13507*, 2025.

Derek Tam, Margaret Li, Prateek Yadav, Rickard Brüel Gabrielsson, Jiacheng Zhu, Kristjan Greenewald, Mikhail Yurochkin, Mohit Bansal, Colin Raffel, and Leshem Choshen. Llm merging: Building llms efficiently through merging. In *NeurIPS 2024 Competition Track*, 2024.

Lukas Tuggener, Mohammadreza Amirian, Fernando Benites, Pius von Däniken, Prakhar Gupta, Frank-Peter Schilling, and Thilo Stadelmann. Design patterns for resource-constrained automated deep-learning methods. *AI*, 1(4):510–538, 2020.

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, et al. Efficient large language models: A survey. *arXiv preprint arXiv:2312.03863*, 2023.

Jingsong Wang, Tom Ko, Zhen Xu, Xiawei Guo, Souxiang Liu, Wei-Wei Tu, and Lei Xie. Autospeech 2020: the second automated machine learning challenge for speech classification. *arXiv preprint arXiv:2010.13130*, 2020.

Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Jordan Zaykov, Jose Miguel Hernandez-Lobato, Richard E Turner, Richard G Baraniuk, Craig Barton, Simon Peyton Jones, et al. Results and insights from diagnostic questions: The neurips 2020 education challenge. In *NeurIPS 2020 Competition and Demonstration Track*, pages 191–205. PMLR, 2021.

Laura Weidinger, Deb Raji, Hanna Wallach, Margaret Mitchell, Angelina Wang, Olawale Salaudeen, Rishi Bommasani, Sayash Kapoor, Deep Ganguli, Sanmi Koyejo, et al. Toward an evaluation science for generative ai systems. *arXiv preprint arXiv:2503.05336*, 2025.

Sanghyun Woo, Kwanyong Park, Inkyu Shin, Myungchul Kim, and In So Kweon. Mtmmc: A large-scale real-world multi-modal camera tracking benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22335–22346, 2024.

Zhen Xu, Lanning Wei, Huan Zhao, Rex Ying, Quanming Yao, Wei-Wei Tu, and Isabelle Guyon. Bridging the gap of autograph between academia and industry: Analyzing autograph challenge at kdd cup 2020. *Frontiers in Artificial Intelligence*, 5:905104, 2022.

Chuanyu Xue et al. KDD Cup 2020 Debiasing Challenge: 6th Place Solution. `https://github.com/ChuanyuXue/KDDCUP-2020`, 2020. Accessed: 2024.

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL `https://openreview.net/forum?id=hizSx8pf0U`.